Hardware Design, Expandability, System Cost and Mean Inter-node Message Distance of Augmented Hypercube Torus and Master-Slave Star-Ring Augmented Hypercube Architectures

Maryam Amiripour, Hamid Abachi
Dept. of Electrical and Computer Systems Engineering
Monash University
Australia
Maryam.Amiripour@eng.monash.edu.au
Hamid.Abachi@eng.monash.edu.au

Abstract— **The need for high computation power, as well as advances in the communications technology have resulted in the rapid development of high-performance message passing architectures. Scalability and cost issues which are part of the performance evaluation in parallel processing systems are recognized to be challenging tasks and are considered as the main measures to identify the suitability of the topology for a given application.**

In this paper the most common message passing architectures together with the Augmented Hypercube Torus (AHT) are described and compared with a newly developed architecture called Master-Slave Star-Ring Augmented Hypercube. Its expandability, hardware cost and Mean Inter-node Message Distance (MIMD) are evaluated for various network sizes and their merits and demerits are highlighted.

Keywords— **Expandability, Routing, Reliable Parallel and Distributed Systems, Performance, Message Passing.**

I. Introduction

Although tightly-coupled systems can provide cost effective improvement on the computing power of a single processor, due to their nature, they may suffer from serious bus contention when global shared memory is used. In an attempt to overcome the limitations of memory contention and rather poor performance associated with shared memory architectures, message-passing systems were introduced. These architectures include Torus, Hypercube and Tree systems.

These message passing systems are mainly used in multi-dimensional configurations. In these topologies, processors instead of having access to a common memory, have their own local memory and communication links to other processors to share information, thereby greatly reducing contention [1].

In general, Hypercube architecture can be expanded by increasing its dimensionality (h). Expanding a Hypercube causes an increase in dimensionality which requires more ports per processor. In general, the maximum number of nodes is limited by the fixed number of processor ports. If each node (N) in a Hypercube architectures is a traditional processor, then it can only communicate with one processor at a time (e.g. over a common bus) [2], [3]. Consequently performance is reduced due to lack of simultaneous communication capability with other nodes. One solution to overcome this issue is to implement the SGI Altix NUMA flex architecture. This product uses an SGI NUMA (cache coherent non-uniform memory access) protocol implemented directly in hardware for performance and a modular packaging scheme. The key to the NUMA flex design of Altix is to use a controller ASIC, referred to as the super Hub (SHub) that interfaces to the Titanium 2 front side bus, together with the memory as well as the I/O subsystem, which further interfaces with other NUMAflex components in the system [4]. This provides simultaneous communication between processors in a true message passing environment. We can implement the proposed architecture by using CR brick which houses 4 NUMA flex nodes totalling 8 Intel Itanium 2 processors. Each NUMA flex node has 12 slots and currently supports 2 GB memory [4]. The CR-brick architecture satisfies our satellite node (slave processor) configuration requirements, which simply means it has eight processors including crossbar switches as routers, so that we can consider each slave component to be equivalent to one CR-brick module.

Proceedings 21st European Conference on Modelling and Simulation
Ivan Zelinka, Zuzana Oplatková, Alessandra Orsoni ©ECMS 2007
ISBN 978-0-9553018-2-7 / ISBN 978-0-9553018-3-4 (CD)

II. A Torus of Augmented Hypercube

Compared to the Hypercube, Torus and Tree networks are infinitely expandable by increasing w (width) or n (levels) and keeping t (dimension) or b (branches) constant. No network re-wiring is needed for the tree when nodes are added to the last level, because nodes are appended onto the unconnected branches of the Tree. For a Torus, only minor network re-wiring is required when nodes are added, because nodes need to be inserted into the network [5].

By connecting Augmented Hypercubes (AHs) in a Torus through Routers, an infinitely expandable network is possible by increasing the torus width. It is also possible to have a Tree of AHs, or to replace the AH with any other "augmented" structure, and substituting these with the nodes in some other structure [6]. For the purposes of comparison in this paper, we will limit our discussion to the AHT, Hypercube, Torus and Tree architectures together with a newly proposed architecture called MSSRAH. Figure 1 illustrates a typical AHT architecture.

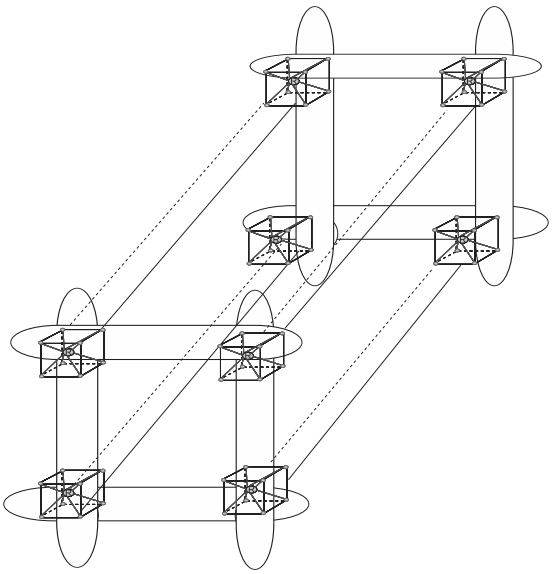

Fig. 1: Augmented Hypercube Torus (AHT) Architecture.

III. Newly Proposed Architecture MSSRAH

We propose a newly developed architecture called Master Slave Star Ring Augmented Hypercube (MSSRAH) configuration. The master processor in this configuration is at the center of the ring and can provide access to each satellite node through fast and reliable communication links. The structure of satellite nodes and the master processor is basically the same although the master processor is faster and has more memory capability plus other supporting hardware and software tools that normally is the requirement for such high speed architecture. This configuration is depicted in Figure 2.

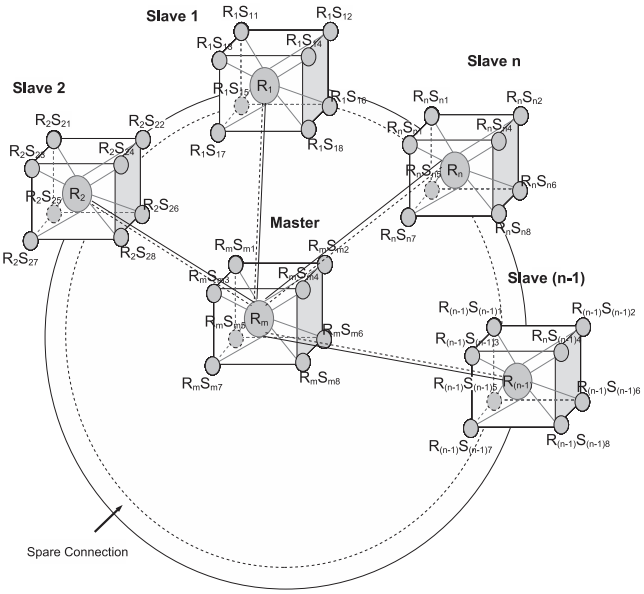

Fig. 2: The Overall Configuration of the Master-Slave Star-Ring Augmented-Hypercube Architecture.

IV. Network Modeling

For a meaningful comparison among popular existing parallel processing architectures with the new ones proposed here, Table 1 summarises parameters of Torus, Hypercube, Tree, AHT and MSSRAH network topologies.

It is straightforward to derive equations for these parameters in the case of the Torus, Hypercube and Tree and AHT networks [7].

The MSSRAH is a star of AHs so that the total number of processing nodes (N_N) is given as the product of the number of Hypercube nodes and the number of Star nodes. This results in:

$$N_N = 2^h \times n \tag{1}$$

where n is the number of Star nodes and 2^h is the number of Hypercube nodes.

The number of communication links (N_L) for the MSSRAH is complicated by the presence of the Router. We will partition the MSSRAH into a Star, Ring and AHs.

TABLE I: Parameters of Torus, Hypercube, Tree, AHT and MSSRAH network topologies.

Architecture	**Number of Nodes(N_N)**	**Number of Links(N_L)**
Torus,t=dim,w=width	w^t	tw^t
Hypercube, h=dim.	2^h	$h2^{h-1}$
Tree,b=branches,n=levels	$\frac{b^n-1}{b-1}$	$\frac{b^n-b}{b-1}$
AHT, t=Torus, w=Torus width, h= Hypercube dim.	$2^h w^t$	$(h2^{h-1}+2^h)w^t + tw^t$
MSSSRAH, n= Number of Star nodes, h= Hypercube dim.	$2^h n$	$(h2^{h-1}+2^h)n + 2(n-1)$

In the MSSRAH there are n AHs, and the number of links in the Star configuration is $n-1$. Since we have a Star topology with n nodes, therefore, the number of links in the Ring configuration is also $(n-1)$. If each AH has N_{LAH} links, then the total number of links is:

$$N_L = 2 \times (n-1) + n \times N_{LAH} \tag{2}$$

Within an AH section there are links directly connected to the processing nodes (in a Hypercube topology), and from each AH node runs a link to the Router. Thus the number of links for a AH segment is:

$$N_{LAH} = h \times 2^{h-1} + 2^h \tag{3}$$

This gives an expression for the number of links for the entire MSSRAH topology, which is dependent on the number of star nodes, and AH dimensionality. This results in:

$$N_L = (h2^{h-1} + 2^h) \times n + 2(n-1) \tag{4}$$

Once the number of processing nodes and the number of links are defined, then the hardware cost analysis can be made as follows.

V. Hardware Cost Analysis

In the context of parallel processing systems, cost is a difficult parameter to define, especially given that component costs are highly dependent on implementation and economic conditions. In general, overall total system cost estimate (C_{ST}) as reported in [8], can be shown as:

$$C_{ST} = C_N(N_N + K_R N_R) + C_L \times N_L \tag{5}$$

where, $K_R = \frac{C_R}{C_N}$, and
- C_N= Cost of node
- C_L= Cost of link
- N_R= Number of Routers
- C_R= Cost of Router

However, a far more difficult question for the multiprocessor system designer is "How well suited is a network over a range of component costs?." In particular one can examine the total system cost when it is compared to the total processing node cost. This can be done since such a figure describes how close a particular network is to the ideal lowest cost network where there are no Router or communication link overhead costs (i.e. $C_R = 0$ and $C_L = 0$, giving $C_{ST} = C_N N_N$) [3]. Thus one can normalise the total system cost function C_{ST}, by $C_N N_N$ giving

$$K_{ST} = \frac{C_{ST}}{C_N N_N} = \frac{K_R N_R + K_L N_L}{N_N} \tag{6}$$

where, $K_L = \frac{C_L}{C_N}$. The normalised total system cost, K_{ST}, then gives us the total system cost relative to the lowest theoretical system cost. The results are summarised in Table 2. It should be noted that only the AHT has a K_R parameter, as it is the only system utilising Routers. For other systems we set $N_R = 0$ as they do not use Routers.

In practice K_L will vary from near zero for a tightly coupled multi-processor system to less than one for a distributed computer network. For tightly and closely coupled multi-processor systems, it has been suggested that $K_L = 0.1$ is a reasonable value [4]. Based on the result of N_L for MSSRAH shown in (4), the K_{ST} can be formulated as follows:

$$K_{ST} = 1 + \frac{K_R N_R + K_L N_L}{2^h n} \tag{7}$$

Since the Number of $N_R = n$ (number of Star nodes), then:

$$K_{ST} = 1 + \frac{K_R}{2^h} + K_L \frac{[(h2^{h-1} + 2^h)n + 2(n-1)]}{2^h n} \tag{8}$$

or,

$$K_{ST} = 1 + K_R 2^{-h} + K_L \frac{[(h2^{h-1}n + 2^h n) + 2n - 2)]}{2^h n} \tag{9}$$

After further simplification and rearrangement, K_{ST} can be expressed as:

$$K_{ST} = 1 + K_R 2^{-h} + K_L [\frac{h}{2} + 1 + \frac{n-1}{2^{h-1}n}] \tag{10}$$

TABLE II: Normalised system cost K_{ST} for Torus, Hypercube, Tree, AHT and MSSRAH architectures.

Parameter	**Torus**	**Hypercube**	**Tree**	**AHT**	**MSSRAH**
K_{ST}	$1+K_L t$	$1+K_L\frac{h}{2}$	$1+K_L\frac{b^n-b}{b^n-1}$	$1+K_L[2^{-h}t+\frac{h}{2}+1]+K_R 2^{-h}$	$1+K_L[1+\frac{h}{2}+\frac{n-1}{n2^{h-1}}]+K_R 2^{-h}$

The overall results for K_{ST} for the message passing architectures including newly proposed architecture are summarised in Table 2.

VI. Significant of K_R for the AHT/ MSSRAH Architecture

Due to the structure of the AHT, the number of Routers is small compared to the number of processing nodes. This may not be true of other structures. It will be demonstrated that for AHT structures of interest in this paper ($h \geq 3$) the cost of Routers is insignificant in comparison to the total system cost [9].

The normalised total system cost for the worst case occurs where links are very cheap ($K_L = 0$). Even with a very pessimistic estimate namely $K_R = 1$ which means the Router's cost is the same as the processing node cost (a realistic system would have a much smaller value, given that processing nodes are far more complex and involve more sub-systems) for a three dimensional AH ($h = 3$), the Routers contribute only %12 to the total system cost. Clearly, from the above formula, the Routers contribute exponentially less as the AH dimensionality increases. Thus, one can ignore the Router's cost contribution and assume $K_R = 0$, whereas the K_{ST} limit for infinite size of MSSRAH would be:

$$1 + K_L[\frac{h}{2} + 1 + \frac{1}{2^{h-1}}] \qquad (11)$$

The graphical presentations of The Normalised System Cost versus the Number of Processing Elements are shown in Figures 3, 4 and 5.

The normalized cost K_{ST} gives us the ratio of the actual total cost to the ideal minimum system cost. In general, communication cost and consequently the system cost increase with increasing K_L. As can be seen in Figures 3, 4 and 5, Torus, AHT and MSSRAH have a constant K_{ST} for different values of h for ($h = 3, 4$ and 6) with $K_L = 0.1$. This implies that communication link costs are always a fraction of the processor costs. Hypercubes differ in this respect because K_{ST} increases as the number of processor nodes increases. This is undesirable since an increasing proportion of the system cost is devoted to communication network overheads and not processors.

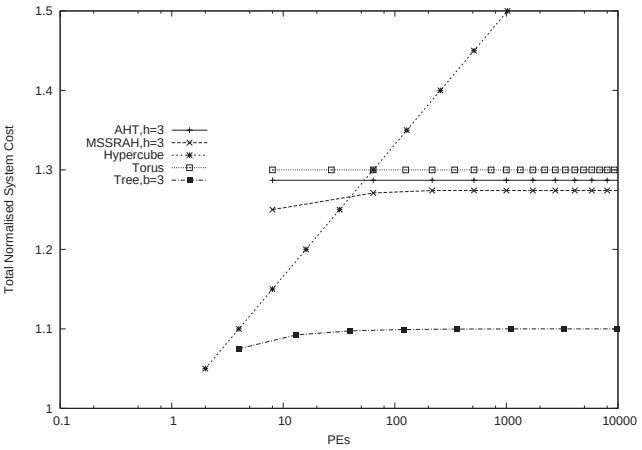

Fig. 3: Normalised System Cost for AHT, MSSRAH, Torus and Tree Networks with h=3 and K_L=0.1.

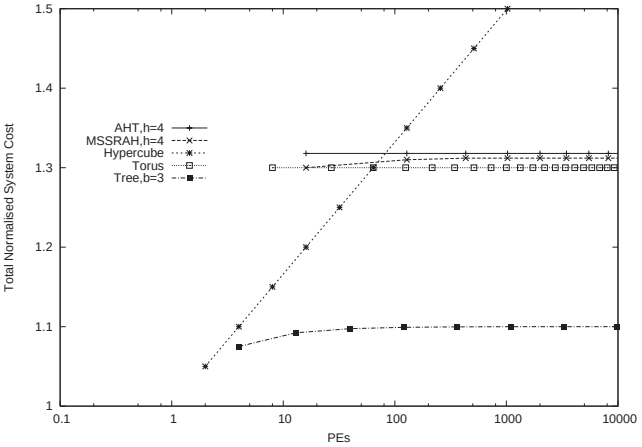

Fig. 4: Normalised System Cost for AHT, MSSRAH, Torus and Tree Networks with h=4 and K_L=0.1.

VII. Mean Inter-node Message Distance (MIMD) Analysis of AHT and MSSRAH Architectures

A. MIMD Analysis of AHT topology

As it is reported in [7], the Mean Inter-node Message Distance (MIMD) under uniformly distributed mes-

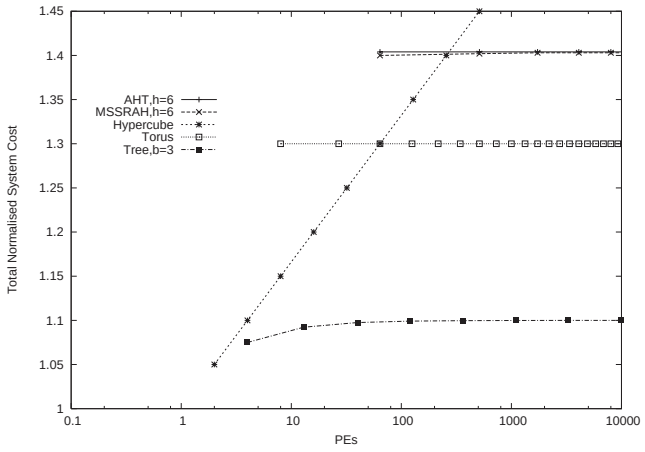

Fig. 5: Normalised System Cost for AHT, MSSRAH, Torus and Tree Networks with h=6 and K_L=0.1.

sages for AHT can be summerised as:

$$MIMD_{AHT} = \frac{tw - tw^{-1} + 8}{4} - \frac{h}{w^t(2^h - 1)}, (w - odd) \quad (12)$$

and,

$$MIMD_{AHT} = \frac{tw + 8}{4} - \frac{h}{w^t(2^h - 1)}, (w - even) \quad (13)$$

B. MIMD Analysis of MSSRAH topology

To find out the average routing distance of MSSRAH, a probabilistic view is taken to identify what fraction of the messages will route within a single AH node, and what fraction of messages will route to different AH nodes.
In this way, the sperate MIMD of AH and the 2-Level Tree can be combined to yield the MIMD of the MSSRAH.
To find the probability that the source (s) and destination (d) nodes are in the same AH (satellite slave) in MSSRAH, we consider the probability of event d, a destination AH node, given that event s, a source AH node has been chosen. The probability that any one particular node is chosen out of n nodes is:
p(s)=p(d)=$\frac{1}{n}$
Based on this assumption, it can be shown as:
$p_{sameAH} = p(d \mid s) = \frac{p(d \cap s)}{p(s)} = \frac{p(d)p(s)}{p(s)}$.
However, since source and destination nodes are independent, therefore:
$p_{SameAH} = p(d) = \frac{1}{n}$
According to [10],

$$MIMD_{AH} = 2 - (\frac{h}{(2^h - 1)})$$

For calculation of Mean Inter-node Message Distance of MSSRAH topology, we consider MSSRAH as a 2-level Tree which its nodes in level 2 have been connected together as a ring. Since all slave Routers have been connected together in pairs, therefore, MIMD for MSSRAH can be shown as:

$MIMD_{MSSRAH} = (MIMD_{2L-Tree_{b=n-1}} + 2) \times (\frac{n-1}{n}) + (MIMD_{AH}) \times (\frac{1}{n})$.
Thus in the general case we have:

$MIMD_{MSSRAH} = [(\frac{2b}{b+1}) + 2] \times (\frac{n-1}{n}) + (2 - \frac{h}{2^h - 1}) \times (\frac{1}{n})$.
After further simplification, rearrangement and considering $b = n - 1$, $MIMD_{MSSRAH}$ can be shown as:

$$MIMD_{MSSRAH} = [\frac{2}{n^2} - \frac{6}{n} + 4] + [2 - \frac{h}{2^h - 1}](\frac{1}{n}) \quad (14)$$

Table 3 provides a summary of different network metrics and Figure 6 illustrates the Mean Inter-node Message Distance for Torus, Tree, Hypercube, AHT and MSSRAH networks.

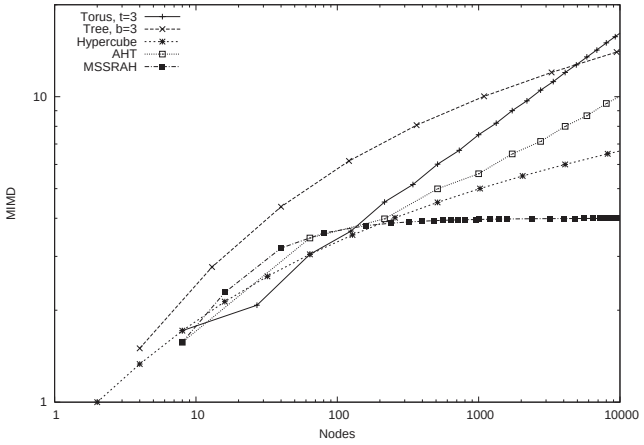

Fig. 6: MIMD for Torus, Tree, Hypercube, AHT and MSSRAH networks.

VIII. Visit Ratio

To calculate the Visit Ratio, we need to know the number of communication links in the network. As reported in [11], the Visit Ratio can be expressed as:

$$VR = \frac{MIMD}{N_L} \quad (15)$$

A. Visit Ratio of AHT topology

As reported in [11],the Visit Ratio of AHT can be shown as:

$$VR_{AHT} = \frac{\frac{tw - tw^{-1} + 8}{4} - \frac{h}{w^t(2^h - 1)}}{(h2^{h-1} + 2^h)w^t + tw^t}, (w - odd) \quad (16)$$

TABLE III: Summary of network metrics

Architecture	**Torus**	**Hypercube**	**Tree**	**AHT**	**MSSRAH**
N_N	w^t	2^h	$\frac{b^n-1}{b-1}$	$2^h w^t$	$2^h n$
diameter	$\frac{tw}{2}$	h	$2(n-1)$	$\frac{tw}{2}+2$	$\frac{n-1}{2}+2$
MIMD	$\frac{tw^{t-1}(w^2-1)}{4(w^t-1)}$ (w-odd) $\frac{tw^{t+1}}{4(w^t-1)}$, (w-even)	$\frac{h2^{h-1}}{2^h-1}$	$\frac{2nb^{n-1}(b^n+1)}{(b^n-1)(b^{n-1}-1)} - \frac{2b^{n-1}(b+1)}{(b^{n-1}-1)(b-1)}$	$\frac{tw-tw^{-1}+8}{4} - \frac{h}{w^t(2^h-1)}$, (w-odd) $\frac{tw+8}{4} - \frac{h}{w^t(2^h-1)}$, (w-even)	$[\frac{2}{n^2}-\frac{6}{n}+4]+[2-\frac{h}{2^h-1}](\frac{1}{n})$
N_L	tw^t	$h2^{h-1}$	$\frac{b^n-1}{b-1}$	$(h2^{h-1}+2^h)w^t+tw^t$	$(h2^{h-1}+2^h)n+2(n-1)$
VR	$\frac{w^2-1}{4w(w^t-1)}$, (w-odd) $\frac{w}{4(w^t-1)}$, (w-even)	$\frac{1}{2^h-13}$	$\frac{2b^{n-2}(b-1)}{b^n-1}$	$\frac{\frac{tw-tw^{-1}+8}{4}-\frac{h}{w^t(2^h-1)}}{(h2^{h-1}+2^h)w^t+tw^t}$, (w- odd) $\frac{\frac{tw+8}{4}-\frac{h}{w^t(2^h-1)}}{(h2^{h-1}+2^h)w^t+tw^t}$, (w-even)	$\frac{[\frac{2}{n^2}-\frac{6}{n}+4]+[2-\frac{h}{2^h-1}](\frac{1}{n})}{[(h2^{h-1}+2^h)n+2(n-1)]n}$

and,

$$VR_{AHT} = \frac{\frac{tw+8}{4} - \frac{h}{w^t(2^h-1)}}{(h2^{h-1}+2^h)w^t+tw^t}, (w-even) \quad (17)$$

B. Visit Ratio of MSSRAH topology

Considering the equation (15), the Visit Ratio of MSSRAH architecture can be evaluated as:

$$VR_{MSSRAH} = \frac{[\frac{2}{n^2}-\frac{6}{n}+4]+[2-\frac{h}{2^h-1}](\frac{1}{n})}{[(h2^{h-1}+2^h)n+2(n-1)]n} \quad (18)$$

IX. Conclusion

This paper examines and compares through mathematical modeling and simulations, the expandability, hardware cost and MIMD analysis of a newly proposed architecture (MSSRAH) with the existing message passing architectures including the AHT, Hypercube, Tree and Torus architectures.

The MSSRAH architecture has a better scalability than the remaining message passing architectures. This simply indicate that it grows with consistency without loss of relative performance. This is evident by having the lower graph for MSSRAH architecture in Figure 6. The MSSRAH architecture not only performs better in terms of the relative cost of other message passing architectures mentioned above, but it also significantly improves the communication performance due to the existence of the Routers and provision of spare Ring link that ensures the system is rarely subject to catastrophic failure.

The distance between the markers on each curve in Figure 6 gives a good indication of the network granularity. Ideally we want the smallest granularity so that the network can be expanded at conveniently small increments.

Careful examination of this feature reveals that the Hypercube has evenly spaced markers which are far apart. This indicates a particular weakness of Hypercube. On the other hand Torus has a desired network granularity. AHT and MSSRAH topologies have an excellent degree of freedom over network granularity in particular MSSRAH, since it has lower curve which is indicative of better performance.

It is also expected that the latter architectures due to the existence of Routers in their centers offer some redundancy in communication so they may become useful in mission critical systems. These topologies require processors with a few communication ports that are supported by SGI technology which would offer performance enhancement.

References

[1] Kodase. S, Wang. S, Gu. Z and Shin. K.G, "Improving scalability of Task Allocation and Scheduling in Large Distributed Real-Time Systems Using Shared Buffers" *Proceedings of the 9th Real-time/Embedded Technology and Applications Symposium (RTAS), IEEE, Washington DC, U.S.A, 2003.*

[2] Walker. J. 1998, " Performance, Reliability and Cost Analysis of Message Passing Architecture" *Master of Engineering Thesis, Department of Electrical and Computer System Engineering, Monash University.*

[3] Bajaj. R, Agrawall. D.J, "Improving Scheduling of Tasks in a Hetergenous Environment" *IEEE Transactions on Parallel and Distributed Systems, Vol 15,No 2,Feb.2004,pp. 107-117.*

[4] Sillicon Graphics Inc. 2004, *Hardware: End-User, Altix 3700 Bx2*, System Overview, Chapter 3, U.S.A, pp.1-6.

[5] Zhuang. X, Libratore. V, "A Recursion-Based Broadcast Paradigm in Wormwhole Routed Networks" *IEEE Transactions on Parallel and Distributed Systems,Vol 16, No 11, Nov. 2005, pp. 1034-1052.*

[6] Abachi. H, Walker. J and Debnath. N. 2000, "Methods for Comparing the Reliability of Advanced Distributed Computer Networks" *International Conference on Computer Applications in Industry and Enginnering, U.S.A, IJCA, pp. 307-310.*

[7] Abachi. H and Walker. J. 1998, "Design and Performance Analysis of Superhypercube, Transputer Tours" *International Journal of Computers and their Applications (ISCA),U.S.A, pp. 1-10.*

[8] Abachi. H and Walker. J. 1996, "Network Expandability and Cost Analysis of Tours, Hypercube and Tree Microproces-

sor Systems" *28th IEEE Southeastern Conference on System Theory, Louisiana, U.S.A, pp. 426-430.*

[9] Abachi. H and Walker. J. 1995, "Scalability and Hardware Cost Analysis of Augmented Hypercube Tours Architecture" *International Conference on Computer Applications in Enginnering and Medicine, Indiana, U.S.A, pp. 232-237.*

[10] Amiripour. M, Abachi. H and Dabke. K. 2007, "Hardware Design, Cost and Diameter Analysis of Super Hypercube Array, Master-Slave Star- Ring Super Hypercube and Master-Slave Super-Super Hypercube 4-Cube architectures" *WSEAS Transactions on Computer Research, to be appeared in 2007.*

[11] Amiripour. M, Abachi. H and Dabke. K. 2007, "Hardware Cost Analysis of Master-Slave Star- Ring Super Hypercube and Master-Slave Super-Super Hypercube 4-Cube Architectures" *6th WSEAS International Conference on Software Engineering Parallel and Distributed Systems (SEPADS'07), Corfu Island, Greece, accepted and to be presented in Feb. 2007.*